ABSTRACT OF THE DISCLOSURE

A method of processing work units in a system is achieved using a throttling queue. The throttling queue assigns work units to a predetermined number of processing slots. Each slot processes an assigned work unit, sends the work units to client systems for processing, and upon finishing, works on a subsequent work unit. The predetermined number of slots, in one embodiment, may be scaled as resources are added or removed from the system. In one embodiment, the work units are network data packets.